THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

On the Stability of Higher Order Digital Filters Which Use Saturation Arithmetic

By J. E. MAZO

(Manuscript received July 15, 1977)

The device of using "saturation arithmetic" to cope with adder overflow in recursive digital filters has, for a number of years now, been known to yield stable operation when the filter is of second order and is linearly stable. Mitra has recently given examples to show that this happy situation does not prevail for higher order filters. Here we investigate conditions on the filter coefficients which would guarantee stability for higher order filters using saturation arithmetic. We are only able to give sufficient conditions for stability. These conditions in their simplest form can be written as linear inequalities involving the coefficients of the filter.

I. INTRODUCTION AND SUMMARY

We shall be concerned with real nth order nonlinear difference equations of the form

$$y(k+n) = f\left[\sum_{i=1}^{n} a_i y(k+n-i)\right], \quad k = 0, 1, 2, \ldots \tag{1}$$

The variables $y(\cdot)$ and the coefficients a_i are real. The initial conditions $y(j)$, $j = 0, 1, \ldots, (n-1)$ are arbitrary, subject only to the important condition $|y(j)| \leq 1$. The function $f(\cdot)$ will be assumed here to have the form given in eq. (2):

$$\begin{aligned} f(x) &= x, \ |x| \leq 1 \\ f(x) &= \operatorname{sgn} x, \ |x| > 1 \end{aligned} \tag{2}$$

This function models a method of handling overflow in the practical implementation of digital filters and in that literature is referred to as "saturation arithmetic." An important unsolved problem is the asymptotic stability of this undriven system. Specifically we would like

to describe the region in "a-space" (i.e. $\{a_i\}$, $i = 1, \ldots, n$) for which

$$\lim_{n \to \infty} y_n = 0$$

for any initial conditions. We always assume that the a_i's are already restricted so that linear stability holds. That is, if $f(\cdot)$ were replaced by the identity function the system would be stable. This is equivalent to all roots λ_i of the characteristic equation

$$c(z) \equiv z^n - \sum_{i=1}^{n} a_i z^{n-i} = 0 \tag{3}$$

satisfying $|\lambda_i| < 1$.* Since $\{a_i\}$ are real, the complex λ_i occur in conjugate pairs.

The case $n = 2$ has special importance to a certain strategy for implementing digital filters and has been considered earlier.[1-3] It was shown for this case that eq. (1) is stable for saturation arithmetic whenever the system is linearly stable. Recently Mitra[4] has shown by example that a result of this generality does not hold for any $n > 2$. This surprising development has regenerated the author's interest in the problem in its own right. In addition, direct implementation of digital filters of the form [eq. (1)] for $n > 2$ is now of interest, so questions of stability must be answered. Saturation arithmetic seems to be an experimentally favored procedure at the moment.

Our main results, Theorems I and II, provide sufficient conditions that a given set of coefficients $\{a_i\}_1^n$ yield a stable filter with saturation arithmetic. Both theorems require one to test if a pair of linear inequalities in a set of variables can be satisfied when the latter lie in a hypercube of (at most) dimension n. A finite algorithm which is sufficient to decide this question is given in Section IV. While we feel that Theorem II will give more powerful results (i.e., determine a larger stability region), Theorem I allows one to list a set of linear inequalities in the a_i, which, if any one is satisfied, would guarantee stability. This result is given as Corollary I.

II. SOME LINEAR AND NONLINEAR THEORY

If one were concerned with the linear version of eq. (1) [$f(\cdot)$ equal to the identity function], the solutions would be

$$y(k) = \sum_{i=1}^{n} k_i \lambda_i^k \quad k = 0, 1, \ldots \tag{4}$$

* Under this condition, stability of eq. (1) with $|y_i| \leq 1$ and

$$\sum_{1}^{m} |a_i| < 1$$

is trivial, since the system is then linear.

where the λ_i are the characteristic roots of eq. (3), assumed to be distinct in writing the solution eq. (4). The constants k_i are determined by initial conditions, or equivalently, the values initially stored in the registers. Suppose we wish to solve for the k_i in terms of $y(0), \ldots, y(n-1)$. This clearly requires the inversion of the matrix

$$V(\lambda) = \begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 & & \lambda_n \\ \lambda_1^2 & & \lambda_n^2 \\ \vdots & & \vdots \\ \lambda_1^{n-1} & & \lambda_n^{n-1} \end{bmatrix}$$

called a Vandermonde matrix. A brief discussion of these matrices is given in the Appendix, as well as some notation we shall use related to them.

The linear difference equation can also be written in matrix form if we introduce the n-vector†

$$Y(k) = \begin{pmatrix} y(k) \\ y(k+1) \\ \vdots \\ y(k+n-1) \end{pmatrix}, \quad k = 0, 1, 2, \ldots \tag{5}$$

or, in component form $y_i(k) = y(k+i-1)$, $i = 1, \ldots, n$. The "time" argument is indicated by the discrete index k. The equation-of-motion is then

$$Y(k+1) = AY(k) \tag{6}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & \cdots & 0 & 1 \\ a_n & a_{n-1} & \cdots & a_2 & a_1 \end{bmatrix} \tag{7}$$

Expanding the determinant of the matrix $A - \lambda I$ (I being the identity matrix) by the last row, we obtain (except for a sign) the polynomial eq. (3), and, not surprisingly, the eigenvalues of A are the roots λ_i mentioned earlier.

If these eigenvalues are distinct, a well-known theorem of algebra guarantees that there exists a nonsingular matrix P such that

$$P^{-1}AP = \Lambda \tag{8}$$

where Λ is simply the diagonal matrix of eigenvalues of A.

† Vectors and matrices will be denoted by capital letters.

Theorem: If the eigenvalues of A are distinct, then

$$V^{-1}AV = \Lambda \tag{9}$$

where V is the Vandermonde matrix formed from the roots of the characteristic eq. (3).

The choice $P = V$ is not claimed to be unique.

Proof:

$$\sum_{j=1}^{n} a_{ij}v_{jk} = \sum_{j=1}^{n} a_{ij}\lambda_k^{j-1} = \begin{cases} \lambda_k^i & i \le n-1 \\ \sum_{j=1}^{n} a_{n+1-j}\lambda_k^{j-1} = \lambda_k^n & \text{for } i = n \end{cases}$$

The second line in the last member makes explicit use of the fact that λ_k is a root of the characteristic equation. Thus

$$\sum_{i=1}^{n} v_{\ell i}^{-1}\left(\sum_{j=1}^{n} a_{ij}v_{jk}\right) = \sum_{i=1}^{n} v_{\ell i}^{-1}(\lambda_k v_{ik}) = \lambda_k \delta_{\ell k}$$

as was to be shown.

One reason for wishing to diagonalize A in the linear case is the simple form that the equation of motion takes. If we multiply eq. (6) by V^{-1} and perform the standard trick of inserting $I = VV^{-1}$ after the A in eq. (6), we obtain

$$Z(k+1) = \Lambda Z(k) \tag{10}$$

where

$$Z(k) = V^{-1}Y(k) \tag{11}$$

Since Λ is diagonal, the solution for the ith component of Z is simply

$$z_i(k) = \lambda_i^k z_i(0) \tag{12}$$

Turning now to nonlinear problems, we wish to summarize some results from Liapunov stability theory,[7] without proofs, and without complete generality.*

We are concerned with an autonomous (time independent) nonlinear difference equation

$$Y(k+1) = F[Y(k)] \tag{13}$$

where F is a nonlinear (or linear) vector function of the vector $Y(k)$.

* A. N. Willson was the first to explicitly apply Liapunov theory to the present problem for $n = 2$.[3] He has also attacked other stability questions for $n = 2$ with these methods in Ref. 8.

If

$$\lim_{k \to \infty} Y(k) = 0$$

then the system is called asymptotically stable.

Theorem (Liapunov): If there exists a (strictly) positive definite quadratic form w[Y], such that, for any allowed n-vector Y, $w[F(Y)] - w[Y] < 0$ (strict inequality) then the system is asymptotically stable.†

In other words, if we can find a positive definite quadratic form (of the state variables of the system) which is always decreasing as the motion proceeds, then the motion must proceed to the origin. The function $w[\cdot]$ is called a Liapunov function.

In terms of $f(\cdot)$ and A the function $F(\ )$ is determined by

$$[F(Y)]_i = [AY]_i \qquad i = 1, \ldots, (n-1)$$

$$[F(Y)]_n = \begin{cases} [AY]_n, & \text{if } |[AY]_n| \leq 1 \\ \operatorname{sgn}[AY]_n, & \text{otherwise} \end{cases} \tag{14}$$

As a simple example of a Liapunov function consider the linear case with nondegenerate eigenvalues, and again set $Z = V^{-1}Y$. Choose

$$w = \sum_{i=1}^{n} |z_i|^2 \tag{15}$$

the z_i being regarded as functions of the $y(j)$, $j = 0, \ldots, n-1$. We know that when $Y \to AY$ we have $Z \to \Lambda Z$ and so

$$w \to \sum_{i=1}^{n} |\lambda_i|^2 |z_i|^2$$

Since we assume $|\lambda_i| < 1$, strict decrease of w is assured.

III. A SPECIAL LIAPUNOV FUNCTION

For the nonlinear problem eq. (1), we shall, for a first pass, choose a $w[\cdot]$ whose form is inspired by the one just described. Noting from the Appendix

$$[V_M^{-1}Y]_i = \frac{(-1)^i \upsilon^{(i)}(\lambda)}{\upsilon(\lambda)} \sum_{y=1}^{n} (-1)^j p_{n-j}^{(i)}(\lambda) y_j$$

$$= \frac{(-1)^{i+n} \upsilon^{(i)}(\lambda)}{\upsilon(\lambda)} \sum_{\ell=0}^{n-1} (-1)^\ell p_\ell^{(i)}(\lambda) y(n-\ell-1) \tag{16}$$

we shall single out the functionals

† For our problem any vector Y is allowed that has components $|y_i| \leq 1$. We identify Y with $Y(0)$ and so $y_i = y(i-1)$, $i = 1, \ldots, n$.

$$x_i = \sum_{\ell=0}^{n-1} (-1)^\ell p_\ell^{(i)}(\lambda) y_{n-\ell} = \sum_{\ell=0}^{n-1} (-1)^\ell p_\ell^{(i)}(\lambda) y(n-\ell-1) \quad (17)$$

for special attention. Clearly the equation of motion for the Z variables (10) implies that under the substitution $Y \to AY$ we have $x_i \to \lambda_i x_i$. We choose

$$w[Y] = \sum_{i=1}^{n} |x_i|^2 \quad (18)$$

where the x_i, via eq. (17), are regarded as functions of y_i, $i = 1, \ldots, n$, the components of Y. Of course we always have $|y_i| \leq 1$.

In order to investigate the consequences of the (sufficient) stability condition

$$w[F(Y)] - w[Y] < 0 \quad (19)$$

we note that under $Y \to AY$ we have $x_i \to x_i^{(L)}$ (L stands for linear) where

$$x_i^{(L)} = \sum_{\ell=0}^{n-1} (-1)^\ell p_\ell^{(i)}(\lambda) y(n-\ell) \quad (20)$$

and

$$y(n) = \sum_{\ell=1}^{n} a_\ell y(n-\ell) \quad (21)$$

Finally the nonlinear (NL) "version" of eq. (20) is

$$x_i^{(NL)} = \begin{cases} x_i^{(L)} \text{ if } |y(n)| \leq 1 \\ \operatorname{sgn} y(n) + \sum_{1}^{n-1} (-1)^\ell p_\ell^{(i)}(\lambda) y(n-\ell) \qquad \text{if } |y(n)| > 1 \end{cases} \quad (22)$$

where we made use of the definition of $F(\cdot)$, and the fact that $p_0^{(\ell)}(\cdot) = 1$. Then

$$w[F(Y)] \equiv \sum_{i=1}^{n} |x_i^{(\mathrm{NL})}|^2 \quad (23)$$

We have already noted that if $|y(n)| \leq 1$ we have a linear iteration $[F(Y) = AY]$ and

$$w[F(Y)] - w[Y] = w[AY] - w[Y]$$

$$= \sum_{i=1}^{n} |\lambda_i|^2 |x_i|^2 - \sum_{i=1}^{n} |x_i|^2 < 0 \quad (24)$$

Hence we will only be concerned with $x_i^{(\mathrm{NL})}$ when $|y(n)| > 1$. In this case,

$w[F(Y)]$ is a function of $y(1), y(2), \ldots, y(n-1)$, while the condition $y(n) > 1$ involves $y(0)$ as well.*

We shall not use the stability condition in the form of eq. (19), but instead note that since

$$\sum_{i=1}^{n} |x_i^{(L)}|^2 = \sum |\gamma_i|^2 |x_i|^2 < \sum |x_i|^2 \tag{25}$$

the condition

$$\sum_{i=1}^{n} |x_i^{(\mathrm{NL})}|^2 - \sum_{i=1}^{n} |x_i^{(L)}|^2 < 0 \text{ when } y(n) > 1 \tag{26}$$

is sufficient for stability. At the price of losing some power in the method we shall see momentarily that we have gained considerably in analytic simplicity. For convenience define

$$c_i = -\sum_{\ell=1}^{n-1} (-1)^{\ell} y(n-\ell) p_{\ell}^{(i)}(\lambda) \tag{27}$$

so that [when $y(n) > 1$]

$$\begin{aligned} x_i^{(L)} &= y(n) - c_i \\ x_i^{(\mathrm{NL})} &= 1 - c_i \end{aligned} \tag{28}$$

Then

$$\begin{aligned} \sum_{1}^{n} |x_i^{(L)}|^2 - \sum_{1}^{n} |x_i^{\mathrm{NL}}|^2 &= \sum_{i=1}^{n} (y(n)-1)(y(n)+1-c_i-c_i^*) \\ &= (y(n)-1)\left[n(y(n)+1) - 2\sum_{1}^{n} c_i\right] \end{aligned} \tag{29}$$

We have used the fact that complex λ_i occur in conjugate pairs and so

$$\sum_i p_{\ell}^{(i)}(\lambda) = \sum_i p_{\ell}^{(i)*}(\lambda)$$

Thus if the inequalities

$$\begin{aligned} 2\sum_{i=1}^{n} c_i - n(1+y(n)) &> 0 \\ y(n) &> 1 \end{aligned} \tag{30}$$

have no solution in the n-cube $|y(i)| \leq 1, i = 0, 1, \ldots, n-1$ the nonlinear equation (1) is stable. More explicitly by using the definition of the c_i [eq. (27)] Lemma I in the Appendix coupled with (67) to express $\Sigma_{i=1}^{n} c_i$ in terms of the a_i, and using finally eq. (21), we have:

* By symmetry of the problem, $|y(n)| > 1$ is here and henceforth replaced with $y(n) > 1$.

Theorem I: If the two inequalities

$$\sum_{\ell=1}^{n} (n - 2\ell) a_\ell y(n - \ell) \geq n \tag{31}$$

$$\sum_{\ell=1}^{n} a_\ell y(n - \ell) > 1, \tag{32}$$

cannot be simultaneously satisfied by some set of $y(i)$ in the cube $|y(i)| \leq 1$, $i = 0, 1, \ldots, n - 1$, then the system (1) is stable for saturation arithmetic.

We note that if the inequalities are satisfied, no conclusion is drawn.

A systematic algorithm for checking the above inequalities is given in the next section. Here we deduce the simple:

Corollary I: If the coefficients a_ℓ satisfy

$$\sum_{\ell=1}^{n} |a_\ell| |n - 2\ell| < n \tag{33}$$

then eq. (1) is stable. If the coefficients a_1 satisfy

$$\sum_{\ell=1}^{n} |a_\ell| |k - \ell| < k \tag{34}$$

for at least one k, $[n/2] < k \leq n$, eq. (1) is stable.*

The first inequality follows immediately from the first inequality in Theorem I plus the fact that $|y(i)| \leq 1$, all i. The remaining inequalities follow from the observation that in eq. (31), the coefficients of a_ℓ have the same sign for $\ell \leq [n/2]$, whereas for $\ell > [n/2]$ they have opposite signs. Thus if eqs. (31–32) have a simultaneous solution, so do

$$\sum_{1 \leq \ell \leq [n/2]} |a_\ell| (n - 2\ell) + \sum_{\ell > [n/2]}^{n} (n - 2\ell) a_\ell y(n - \ell) \geq n \tag{35}$$

$$\sum_{\ell \leq [n/2]} |a_\ell| + \sum_{\ell > [n/2]} a_\ell y(n - \ell) > 1 \tag{36}$$

where the above is obtained by setting $y(n - \ell) = \operatorname{sgn} a_\ell$, $1 \leq \ell \leq [n/2]$. If, for $k > [n/2]$, we multiply the second inequality by $(2k - n) > 0$ and add the result to the first we obtain

$$\sum_{\ell \leq [n/2]} |a_\ell| (k - \ell) + \sum_{\ell > [n/2]} (k - \ell) a_\ell y(n - \ell) \geq k \tag{37}$$

which, if $|y(i)| \leq 1$, cannot possibly be satisfied when

* The notation $[x]$ denotes the integer part of x.

$$\sum_{\ell=1}^{n} |a_\ell| |k - \ell| < k \tag{38}$$

There is one important point to be noted here. Our results have only been proven for the case of nondegenerate eigenvalues. For degenerate eigenvalues, the Liapunov function that we have chosen in this section is not strictly positive definite. We have in fact constructed Liapunov functions which are specifically designed to handle the degenerate case. Using them, we have proven that the conclusions are still true for degenerate eigenvalues. We believe that the form of the results, being simple conditions on the a_l's, will allow the reader to readily accept that they are true in general. Since our proof of the extension is long and out of proportion to its importance, we have chosen to omit it.

IV. HYPERPLANE ALGORITHM

Let $\{b_i\}_{i=1}^k$ and $\{c_i\}_{i=1}^k$, ξ and η denote fixed constants. We wish to determine when it is possible to simultaneously satisfy the inequalities

$$\sum_{i=1}^{k} z_i b_i \geq \xi$$

$$\sum_{i=1}^{k} z_i c_i \geq \eta$$

$$|z_i| \leq 1 \quad i = 1, 2, \ldots, k \tag{39}$$

The dimensionality of the problem is immediately reduced if sgn b_j = sgn c_j for some j since we may immediately take z_j = sgn b_j. It is important to note that we assume this to have been done and therefore assume $b_i c_i < 0$, $1 \leq i \leq k$.

Lemma: If the simultaneous inequalities eq. (39) are satisfied then there exists $\tilde{z}_i$, and a j, $1 \leq j \leq k$, such that

$$\sum_{i=1}^{k} \tilde{z}_i b_i \geq \xi$$

$$\sum_{i=1}^{k} \tilde{z}_i c_i \geq \eta$$

$$|\tilde{z}_j| \leq 1, \; |\tilde{z}_\ell| = 1 \text{ all } \ell \neq j \tag{40}$$

In other words all but possibly one of the coordinates may be given values ± 1.

This is geometrically evident if $k = 2$. If $k > 2$ one need only consider the z_i variables two at a time, always applying the Lemma for $k = 2$. Eventually all but perhaps one of the z_i will have value ± 1.

Continuing with the description of the algorithm, let E be any k-vector

having components $\epsilon_\ell = \pm 1$; there are 2^k such E vectors. Choose a j and let's test if indeed it is the j of the Lemma. Let $\tilde{z}_\ell = \epsilon_\ell$, $\ell \neq j$. Then if eq. (40) has this solution we have

$$\frac{\xi - \Sigma' b_\ell \epsilon_\ell}{b_j} \leq \tilde{z}_j \leq \frac{\eta - \Sigma' \epsilon_\ell c_\ell}{c_j} \quad \text{if } b_j > 0 \tag{41a}$$

or

$$\frac{\eta - \Sigma' \epsilon_\ell c_\ell}{c_j} \leq \tilde{z}_j \leq \frac{\xi - \Sigma' b_\ell \epsilon_\ell}{b_j} \quad \text{if } b_j < 0 \tag{41b}$$

If these inequalities are consistent (i.e., the upper bound is at least as big as the lower bound) and if they can be satisfied by some $\tilde{z}_j$, $|\tilde{z}_j| \leq 1$ we are done—the simultaneous inequalities are satisfied. If not, try another E vector, or another j. For a given j there are 2^{k-1} E vectors to try. Hence after at most $k \cdot 2^{k-1}$ such attempts we have exhausted all things that need to be checked, and checking the inequalities is, it has turned out, a finite procedure.

This procedure is not only applicable to Theorem I, but also to Theorem II occurring in Section V.

V. ANOTHER LIAPUNOV FUNCTION

We have already noted that the entire sequence $\{y_i\}_0^\infty$ is determined by the first n elements. For our second choice of the Liapunov function we chose the expression for the energy in the remainder of the sequence for the linear problem:

$$w[Y] = \sum_n^\infty y_k^2 \quad \text{(linear case)} \tag{42}$$

The right member is regarded as a positive definite quadratic form in $Y(0)$. In the linear case we also have, numerically,

$$w[AY] = \sum_{n+1}^\infty y_k^2 \tag{43}$$

which is smaller than $w[Y]$ by $y^2(n)$. Thus this $w[\cdot]$ doesn't necessarily decrease after every iteration and thus it is not strictly a Liapunov function. However after at most n iterations it must decrease (unless all $y_i = 0$) and the effect will be the same. We shall have stability if we can show whenever $y(n) > 1$, that

$$w[F(Y)] - w[Y] < 0 \tag{44}$$

or, equivalently

$$w[F(Y)] - w[AY] < y_n^2 \tag{45}$$

We begin by writing down the generating function for the sequence $y(n), y(n+1), \ldots$, when linear theory holds. By definition

$$H(z) \equiv \sum_{k=0}^{\infty} y(n+k)z^k \tag{46}$$

Using standard linear techniques we calculate from eq. (1) and its initial conditions

$$H(z) = \frac{\sum_{j=1}^{n} y(n-j) \sum_{s=0}^{n-j} a_{j+s} z^s}{-\sum_{s=0}^{n} a_s z^s} \tag{47}$$

where we have arbitrarily defined $a_0 = -1$. We note the characteristic polynomial

$$c(z) = -\sum_{i=0}^{n} a_i z^{n-i}$$

has the same modulus as the denominator of $H(z)$ when $|z| = 1$ (since the a_i are real).

We next note that

$$\sum_{k=n}^{\infty} y_k^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(z = e^{i\theta})|^2 d\theta \tag{48}$$

To express this as a quadratic form introduce, for $0 \le |s-t| \le (n-1)$ the integrals ($z = e^{i\theta}$)

$$I_{st} \equiv \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{z^{s-t}}{\left| \sum_{s=0}^{n} a_s z^s \right|^2} d\theta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\cos(s-t)\theta}{\left| \sum_{s=0}^{n} a_s z^s \right|^2} d\theta \tag{49}$$

Also, as a contour integral around the unit circle, we have

$$I_{st} = \frac{1}{2\pi i} \oint \frac{z^{n+s-t-1}}{(\Sigma a_s z^s)(\Sigma a_t z^{n-t})} dz \tag{50}$$

The first form of the integrals shows they are real and $I_{st} = I_{ts}$. Also introduce the real symmetric matrix

$$H_{jk} = \sum_{s=0}^{n-j} \sum_{t=0}^{n-k} a_{j+s} a_{k+t} I_{st}, \quad j, k = 1, \ldots, n \tag{51}$$

Note for $j = k$, $H_{jj} > 0$ since the right side of eq. (51) is then a Teoplitz form with positive spectral function. Then with these notations simply using eq. (47) to expand the integral in eq. (48) yields

$$w[Y] = \sum_{k=n}^{\infty} y_k^2 = \sum_{j,k=1}^{n} y(n-j)y(n-k)H_{jk} \tag{52}$$

or

$$w[AY] = \sum_{k=n+1}^{\infty} y_k^2 = \sum_{j,k=1}^{n} y(n+1-j)y(n+1-k)H_{jk} \tag{53}$$

When $y_n > 1$, $w[F(Y)]$ has the same form as eq. (53) except that y_n is replaced by unity. Thus in evaluating $w[F(Y)] - w[AY]$ [by using eq. (53)] most of the quadratic terms will cancel. Doing this and a few minor manipulations, the criterion for stability will read that we must have

$$[y(n)+1]H_{11} + 2\sum_{j=2}^{n} H_{1j}y(n+1-j) \geq -\frac{y^2(n)}{y(n)-1} \tag{54}$$

whenever

$$y(n) = \sum_{j=1}^{n} a_j y(n-j) > 1 \tag{55}$$

If we define $H_{1,n+1} = 0$, we may write this as:

Theorem II: If there are no simultaneous solutions to the inequalities

$$\sum_{i=1}^{n} [-a_i H_{11} - 2H_{1,i+1}]y(n-i) \geq H_{11} + \frac{y^2(n)}{y(n)-1} \tag{56}$$

$$y(n) = \sum_{j=1}^{n} a_j y(n-j) > 1 \tag{57}$$

$$|y(i)| \leq 1 \quad i = 0, 1, \ldots, n-1 \tag{58}$$

then the difference eq. (1) is stable with saturation arithmetic.

To convert this into a hyperplane problem (discussed in Section IV) the nonlinear term $y^2(n)/(y(n)-1)$ may be dropped or replaced by the value four (since

$$\frac{x^2}{x-1} \geq 4$$

when $x \geq 1$). If we drop this term, the resulting stability has the physical interpretation that at any time the energy in the remaining tail of the nonlinear response is less than or equal to the corresponding energy for the linear problem, regardless of the previous state. That is, if measured by energy, the nonlinear undriven response dies off at least as fast as the linear response for any initial conditions.

We also note that whenever the Liapunov function $w = Y^{\dagger}HY$ [eq. (52)] drops to the value unity the system behaves as a linear one from there on (no future y_k will exceed unity). Several bounds for this quantity may be given. Since

$$\left|\sum_{s=0}^{\infty} a_s z^s\right|^2 = \prod_{i=1}^{n} |1 - z\lambda_i|^2 \geq \prod_{i=1}^{n} (1 - |\lambda_i|)^2$$

we have

$$w = \frac{1}{2\pi}\int_{-\pi}^{\infty} d\theta |H(z)|^2 \le \max_{z=e^{i\theta}} |H(z)|^2$$

$$= \max_{z=e^{i\theta}} \left| \frac{\sum_{j=1}^{n} y(n-j) \sum_{s=0}^{n-j} a_{j+s} z^s}{-\sum_{0}^{n} a_s z^s} \right|^2$$

$$\le \frac{\left| \sum_{j=1}^{n} |y(n-j)| \sum_{s=0}^{n-j} |a_{j+s}| \right|^2}{\prod_{i=1}^{n} (1 - |\lambda_i|)^2} \le \frac{\left| \sum_{j=1}^{n} |a_j| \right|^2 \left| \sum_{i=0}^{n-1} |y(i)| \right|^2}{\prod_{i=1}^{n} (1 - |\lambda_i|)^2} \tag{59}$$

VI. EXAMPLES

The few simple examples of this section will shed light on the two methods we have given. When $n = 2$ the two inequalities of Corollary I are simply $|a_2| < 1$, $|a_1| < 2$. Since linear stability implies $|a_1| = |\lambda_1 + \lambda_2| < 2$, $|a_2| = |\lambda_1\lambda_2| < 1$, we see that for $n = 2$ linear stability implies stability with saturation arithmetic. We have already mentioned this is not true for $n > 2$. Mitra has constructed a counter example using the degenerate case $\lambda_1 = \lambda_2 = \ldots = \lambda_n \equiv \gamma$. For $n = 3$ he finds oscillations if $|\gamma| \ge 0.858$, although if $|\gamma|$ is smaller than this stability is not implied. If we consider the second inequality of Corollary I for $n = 3, k = 2, a_1 = 3\gamma, a_2 = -3\gamma^2, a_3 = \gamma^3$ we have stability if

$$3|\gamma| + |\gamma|^3 < 2$$

or $|\gamma| < 0.596$. No better result is obtained for this case by a complete use of Theorem I.

On the other hand if we apply the criterion of Theorem II (neglecting the nonlinear term) with the algorithm of Section IV, we find stability if $|\gamma| < 0.71$. Insignificant improvement would be obtained here if we had also included the nonlinear term. The application of Theorem II to the present case was sufficiently simple so that the calculation could be done by hand. The integrals were done exactly to give

$$H_{11} = \frac{\lambda^2}{(1-\lambda^2)^5}[9 - 9\lambda^2 + 10\lambda^4 - 5\lambda^6 + \lambda^8]$$

$$H_{12} = \frac{-3\lambda^3}{(1-\lambda^2)^5}(3 + \lambda^2)$$

$$H_{13} = \frac{3\lambda^4}{(1-\lambda^2)^5}(1 + \lambda^2) \tag{60}$$

Clearly in general the integrals would have to be done numerically. The fact that the nonlinear term did not contribute significantly is due to the combination of facts that at the critical value for λ, $y(n = 3)$ is not extremely close to one and also H_{11} is large, due to its denominator.

It should be pointed out that

$$\sum_{n}^{\infty} y_k^2$$

can be expressed using the solution eq. (4) for the linear problem as

$$\sum_{n}^{\infty} y_k^2 = Y^+(V^{-1})^\dagger \Gamma V^{-1} Y \tag{61}$$

where

$$\Gamma_{ij} = \frac{(\gamma_i^*)^n \lambda_j^n}{1 - \gamma_i^* \gamma_j} \tag{62}$$

However this explicit form is only true for the nondegenerate case and, although all limits exist as $\lambda_1 \to \lambda_2$, etc., the expression would probably not be suitable for numerical computation when eigenvalues are close to being degenerate.

Another example of limiting misbehavior being only apparent is that in a similar manner one could compute that

$$H(z) = \sum_{i=1}^{n} \frac{1}{1 - z\lambda_i} (V^{-1}Y)_i \tag{63}$$

Individual terms in this expression are badly behaved as, for example, if $\lambda_1 \approx \lambda_2$, but the alternate form eq. (47) shows everything is well behaved in the limit.

If we return to Corollary I applied to $n = 3$ when $(\lambda_1, \lambda_2, \lambda_3) = \rho(i, -i, 1)$, $0 < \rho < 1$, we see that the inequality

$$|a_1| + |a_3| < 2$$

is sufficient to guarantee that for filter poles in these relative position saturation arithmetic will give a stable filter for any ρ, all the way out to the boundary of linear stability.

Based on these examples we feel that Theorem II is the more powerful method although the simpler Corollary I can yield considerable information for particular cases.

Finally, we note that an important investigation on the present problem has just been completed by Mitra,[9] resulting in different stability criteria from those presented here. Mitra's results will give a polynomial type criterion for absence of *periodic* oscillations. These results in themselves do not prove stability in that Ref. 9 does not preclude unending periodic outputs with no input. However, we take the

liberty of mentioning that Mitra has extended the proof to include stability and thus the results of Ref. 9 may be taken as proving the same type of stability as discussed here. Another comparison with Ref. 9 involves the size of the stability region in "tap-space" which the two methods give. Neither Mitra's criterion nor ours can claim to describe the largest stability region. Also it does not even seem possible at this stage to give theoretical arguments to decide if one of the methods is always superior in this respect. However, several examples indicate that the region determined by Mitra's criterion is larger. Assuming this to be the case in general, an effective practical procedure would be to first test for stability using our simple Corollary I, and if this fails, apply Mitra's polynomial test.

APPENDIX

The Vandermonde Matrix and Symmetric Polynomials

The α denote an ordered set of n complex members α_i, i.e., $\alpha_1, \alpha_2, \ldots, \alpha_n$. By the Vandermonde matrix $V(\alpha)$, we shall mean the matrix

$$V(\alpha) = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \alpha_3 & \ldots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \alpha_3^2 & \ldots & \alpha_n^2 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \alpha_3^{n-1} & \ldots & \alpha_n^{n-1} \end{bmatrix} \tag{64}$$

This can be written $[V(\alpha)]_{ij} = \alpha_j^{i-1}$, $i, j = 1, \ldots, n$. We let $v(\alpha) = \det V(\alpha)$, and it is known[5] that

$$v(\alpha) = \Pi(\alpha_j - \alpha_i) \tag{65}$$

where the product extends over all i, j satisfying

$$1 \leq i < j \leq n$$

If $\alpha_i \neq \alpha_j$ for $i \neq j$ then the inverse of $V(\alpha)$ exists, and is known. Before giving its structure, we wish to list some facts concerning some special symmetric polynomials.[6]

Definition: The ℓth elementary symmetric function of n-variables ($\ell = 1, 2, \ldots, n$) is the sum of all formally distinct products of the variables taken ℓ at a time. We also define $p_0 \equiv 1$.

For example if $n = 3$ we have

$$p_0(\alpha) = 1$$

$$p_1(\alpha) = \alpha_1 + \alpha_2 + \alpha_3$$

$$p_2(\alpha) = \alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3$$

$$p_3(\alpha) = \alpha_1\alpha_2\alpha_3 \tag{66}$$

A well known theorem of algebra states that any symmetric polynomial in all the α's can be uniquely written as a *polynomial* in the quantities $p_i(\alpha)$, $i = 0, \ldots, n$.

Note that in the characteristic polynomial eq. (3) we have

$$a_i = (-1)^{i+1}p_i(\lambda) \tag{67}$$

where the λ_i, $i = 1, \ldots, n$, are the roots of eq. (3). Thus the theorem just stated says that any symmetric polynomial in the roots of a polynomial can be expressed as a polynomial in the coefficients of the equation (rather than a complicated function as would be required to express an individual root).

We shall use the notation $p_\ell^{(i)}(\alpha)$, $\ell = 0, \ldots, n-1$, to denote the ℓth elementary symmetric function formed from the $(n-1)$ ordered variables $\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_n$. Likewise $v^{(i)}(\alpha)$ denotes the determinant of the corresponding $(n-1) \times (n-1)$ Vandermonde matrix.

Theorem:

$$[V^{-1}(\alpha)]_{ij} = \frac{(-1)^{i+j}v^{(i)}(\alpha)}{v(\alpha)}\, p_{n-j}^{(i)}(\alpha) \tag{68}$$

$$i, j = 1, \ldots, n$$

Proof: We have

$$q_{kj} \triangleq \sum_{i=1}^{n} [V(\alpha)]_{ki}[V^{-1}(\alpha)]_{ij} = \sum_{i=1}^{n} \alpha_i^{k-1} \frac{(-1)^{i+j}v^{(i)}(\alpha)p_{n-j}^{(i)}(\alpha)}{v(\alpha)} \tag{69}$$

From eq. (65) and the definition of $v^{(i)}(\alpha)$ we see that

$$\frac{v^{(i)}(\alpha)}{v(\alpha)} = \frac{1}{(-1)^{n-1}\prod_{\ell \neq i} (\alpha_i - \alpha_\ell)} \tag{70}$$

and hence eq. (69) is

$$q_{kj} = \sum_{i=1}^{n} \frac{\alpha_i^{k-1}p_{n-j}^{(i)}(\alpha)(-1)^{n-j}}{\prod_{\ell \neq i} (\alpha_i - \alpha_\ell)} \tag{71}$$

Form

$$q(x) \triangleq \sum_{j=1}^{n} q_{kj} x^j$$

and note that*

$$\sum_{j=1}^{n} p_{n-j}^{(i)}(\alpha)(-1)^{n-j} x^j = \sum_{\ell \neq i} (x - \alpha_\ell) \tag{72}$$

Thus

$$\sum_{j=1}^{n} q_{kj} x^{j-1} = \sum_{i=1}^{n} \alpha_i^{k-1} \frac{\prod_{\ell \neq i} (x - \alpha_\ell)}{\prod_{\ell \neq i} (\alpha_i - \alpha_\ell)} \tag{73}$$

has value α_m^{k-1} when $x = \alpha_m$, $m = 1, \ldots, n$. From this it follows that $q(x) = x^{k-1}$, so that $q_{kj} = \delta_{kj}$, which we were to prove.

We leave it to the reader to convince himself of the following:

Lemma I:

$$\sum_{\ell=1}^{n} p_{n-j}^{(\ell)}(\alpha) = j p_{n-j}(\alpha) \quad j > 0 \tag{74}$$

Lemma II:

$$p_j(\lambda) - p_j^{(i)}(\lambda) = \lambda_i p_{j-1}^{(i)}(\lambda) \quad j = 1, 2, \ldots, n-1 \tag{75}$$

REFERENCES

1. P. M. Ebert, J. E. Mazo, and M. G. Taylor, "Overflow Oscillations in Digital Filters," B.S.T.J. *48*, No. 9 (November 1969), pp. 2999–3020.
2. I. W. Sandberg, "A Theorem Concerning Limit Cycles in Digital Filters," Proc. 7th Annual Allerton Conf. Circuits and Systems Theory, pp. 63–68, 1969.
3. A. N. Willson, Jr., "Limit Cycles Due to Adder Overflow in Digital Filters," IEEE Trans. Circuit Theory, *CT-19*, No. 4, pp. 342–346, 1972.
4. D. Mitra, "Large Amplitude, Self-Sustained Oscillations in Difference Equations Which Describe Digital Filter Sections Using Saturation Arithmetic," IEEE Trans. Acoustics, Speech and Signal Processing, *ASSP-25* (1977), No. 2, pp. 134–143.
5. R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, 2nd ed., 1970, p. 193.
6. M. Bocher, *Introduction to Higher Algebra*, Macmillan, 1936, Chapter 18.
7. J. LaSalle and S. Lefschetz, *Stability by Liapunov's Direct Method*, Academic Press, 1961.
8. A. N. Willson, Jr., "Some Effects of Quantization and Adder Overflow on the Forced Response of Digital Filters," B.S.T.J. *51*, No. 4 (April 1972), pp. 863–867.
9. D. Mitra, "Criteria for Determining if a High Order Filter using Saturation Arithmetic is Free of Overflow Oscillations," B.S.T.J., *56*, No. 9 (November 1977), pp. 1679–1700.

* The right member of (72) is a polynomial expressed directly in terms of roots, the left member, via (67), the polynomial expressed directly in terms of coefficients.